ON THE TOTAL VARIATION DISTANCE BETWEEN THE BINOMIAL RANDOM GRAPH AND THE RANDOM INTERSECTION GRAPH

JEONG HAN KIM, SANG JUNE LEE, AND JOOHAN NA

Abstract. When each vertex is assigned a set, the intersection graph generated by the sets is the graph in which two distinct vertices are joined by an edge if and only if their assigned sets have a nonempty intersection. An interval graph is an intersection graph generated by intervals in the real line. A chordal graph can be considered as an intersection graph generated by subtrees of a tree. In 1999, Karoński, Scheinerman and Singer-Cohen [Combin Probab Comput 8 (1999), 131–159] introduced a random intersection graph by taking randomly assigned sets. The random intersection graph $G(n, m; p)$ has n vertices and sets assigned to the vertices are chosen to be i.i.d. random subsets of a fixed set M of size m where each element of M belongs to each random subset with probability p, independently of all other elements in M. Fill, Scheinerman and Singer-Cohen [Random Struct Algorithms 16 (2000), 156–176] showed that the total variation distance between the random graph $G(n, m; p)$ and the Erdös-Rényi graph $G(n, \hat{p})$ tends to 0 for any $0 \le p = p(n) \le 1$ if $m = n^\alpha$, $\alpha > 6$, where $\hat{p}$ is chosen so that the expected numbers of edges in the two graphs are the same. In this paper, it is proved that the total variation distance still tends to 0 for any $0 \le p = p(n) \le 1$ whenever $m \gg n^4$.

1. Introduction

The *intersection graph* on $V := \{1, \ldots, n\}$ generated by a collection $\{L_1, \ldots, L_n\}$ of sets is the graph on V in which two distinct vertices i and j are adjacent if and only if their corresponding sets L_i and L_j have a nonempty intersection. In 1945, Szpilrajn-Marczewski [29] observed that every graph may be represented as an intersection graph. Later, Erdős, Goodman and Pósa [12] showed that every graph with n vertices can be represented as an intersection graph generated by subsets of a set of $n^2/4$ elements. An interval graph is an intersection graph generated by intervals in the real line. A chordal graph turned out to be an intersection graph generated by subtrees of a tree [14]. In general, a class of graphs is called an *intersection class* of a family $\mathcal{F}$ of sets if each graph in the class is an intersection graph generated by sets in $\mathcal{F}$. Scheinerman [27] found a necessary and sufficient condition for a class of graphs to be an intersection class of a family $\mathcal{F}$ of sets. Intersection graphs have been applied to phylogeny problems in biology [17], seriation problems in psychology [18], and contingency tables in statistics [21], etc. For more details, see [24].

In 1999, Karoński, Scheinerman and Singer-Cohen [20] introduced the *random intersection graph*, which is the intersection graph generated by independent and identically distributed (i.i.d.) random subsets $L_1, ..., L_n$ of $M = \{1, ..., m\}$. Fill, Scheinerman and Singer-Cohen [13] considered conditions under which the random intersection graph is essentially the binomial random graph (that is, the Erdős-Rényi random graph with independently chosen edges) with the same expected number of edges. Let $G(n, m; p)$ denote the random intersection graph generated by i.i.d. random subsets $L_1, ..., L_n$ whose distributions are binomial with parameters (m, p), i.e., for a subset A of M, $\Pr[L_i = A] = p^{|A|}(1 - p)^{m-|A|}$. Fill, Scheinerman and Singer-Cohen were interested in how close $G(n, m; p)$

Date: February 13, 2017.

The first author was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) (NRF-2012R1A2A2A01018585) and KIAS internal Research Fund CG046001.

The second author was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A1A1059913). This work was partially done while the second author was visiting Korea Institute for Advanced Study (KIAS).

The third author was supported by KIAS internal Research Fund CG053601.

is to $G(n,\hat{p})$ in terms of total variation distance, where $\hat{p}$ is chosen so that the expected numbers of edges in the two graphs are the same, i.e.,

$$\hat{p} := 1-(1-p^2)^m.$$

The total variation distance between two (graph-valued) random variables X and Y is defined by

$$\mathrm{TV}\,(X,Y) = \frac{1}{2}\sum_G \Big| \Pr[X=G] - \Pr[Y=G] \Big|,$$

where the sum is taken over all possible values of X and Y.

Theorem 1.1 ([13, Theorem 10])**.** *Let $\alpha > 6$ be a constant and $m = n^\alpha$. Then for any $0 \le p = p(n) \le 1$,*

$$\mathrm{TV}\Big(G(n,m;p), G(n,\hat{p})\Big) = o(1).$$

For $3 < \alpha \le 6$, Rybarczyk [26] proved a weaker result. Namely, for any monotone property $\mathcal{A}$, $\Pr[G(n,m;p) \in \mathcal{A}]$ and $\Pr[G(n,\hat{p}) \in \mathcal{A}]$ are essentially the same. The exact statements of the theorems there are rather complicated.

A random intersection graph has received a lot of attention due to a great diversity of applications in areas such as epidemics [9], circuit design [20], network user profiling [23] and analysis of complex networks [3, 4, 7, 10]. For more information, we refer the reader to the survey papers [5, 6, 28]. For instance, $G(n,m;p)$ is applicable for gate matrix circuit design, which is related to the optimization problem of finding a permutation of the order of gate lines that minimizes the number of horizontal tracks required to lay out the circuit. The problem is NP-hard in general, but it is solvable in $O(n)$ time when G is an interval graph [16]. Karoński, Scheinerman and Singer-Cohen [20] studied conditions for which $G(n,m;p)$ is an interval graph with high probability.

When L_i's are uniformly distributed in the class of subsets of M of the same size, the random intersection graph generated by the L_i's is called a *uniform random intersection graph*. An application to security of wireless sensor networks [2, 8, 11, 25] is one of the main motivations for studying the uniform random intersection graph. The random intersection graph can be generalized in the way that the vertices i and j are adjacent if L_i and L_j have at least $s \ge 1$ common elements. The generalization is applicable for cluster analysis [4, 7, 15].

The random intersection graph $G(n,m;p)$ may be defined using an $n \times m$ random matrix $R(n,m;p)$ whose rows are indexed by $i \in V$ and columns are indexed by $a \in M$. Each entry of the matrix is 1 or 0 with probability of p and $1-p$, respectively, independently of all other entries. The row vector indexed by $i \in V$ corresponds to the subset L_i of M. On the other hand, the column vector indexed by $a \in M$ corresponds to the set V_a of all vertices $i \in V$ with $a \in L_i$. The graph $G(n,m;p)$ may be alternatively constructed by taking the edge set to be the union of edge sets of the complete graphs on V_a for all $a \in M$.

The main difference between $G(n,m;p)$ and $G(n,\hat{p})$ are the complete graphs induced by the column vectors with three or more 1's. In particular, the triangles formed by the columns with exactly three 1's play an important role. Those triangles are to be called *artifact triangles*. Roughly speaking, if mp^2 is large, then $\hat{p}$ is close to 1 so that both of $G(n,m;p)$ and $G(n,\hat{p})$ are almost the complete graphs with high probability. On the other hand, if mp^2 is small, then the expected number of artifact triangles is $\binom{n}{3}mp^3(1-p)^{n-3} = O(\frac{n^3}{m^{1/2}}(mp^2)^{3/2})$, which goes to 0, provided $m \gg n^6$. Theorem 1.1 was proved based upon this observation.

In this paper, we will show that the total variation distance is still small enough even if there are some artifact triangles. It is actually small as long as the expected number of pairs of distinct artifact triangles with a common edge is small. If the expected number is not small, the total variation distance may be small when both of $G(n,m;p)$ and $G(n,\hat{p})$ are almost the complete graphs with high probability. Based on these two facts, we infer that if $m \gg n^4$ then the total variation distance is always small for any p: It turns out that the expected number is $O(n^4m^2p^6)$. To have the total

variation distance small for all p, it is required that mp^2 is large when $n^4m^2p^6 = \frac{n^4}{m}(mp^2)^3$ is not small, which holds if $m \gg n^4$.

Theorem 1.2. *For $m \gg n^4$ and $0 \le p = p(n) \le 1$, we have that*

$$\mathrm{TV}\Big(G(n,m;p), G(n,\hat{p})\Big) = o(1).$$

In the next section, we give the outline of the proof of Theorem 1.2. The proof will be divided into four parts, which will be proved in Sections 2-5.

2. Preliminaries and Outline of proof of Theorem 1.2

If $p \ge \left(\frac{3\log n}{m}\right)^{1/2}$, both of $G(n,m;p)$ and $G(n,\hat{p})$ are the complete graphs with probability $1-O\left(\frac{1}{n}\right)$. Indeed, for each edge e,

$$\Pr[e \notin G(n,m;p)] = (1-p^2)^m \le e^{-mp^2} \le \frac{1}{n^3},$$

and hence $G(n,m;p)$ is the complete graph with probability $1-O\left(\frac{1}{n}\right)$. Since the expected numbers of edges in $G(n,m;p)$ and $G(n,\hat{p})$ are the same, $G(n,\hat{p})$ is the complete graph with probability $1-O\left(\frac{1}{n}\right)$ as well. Therefore,

$$\mathrm{TV}\Big(G(n,m;p), G(n,\hat{p})\Big) = O\Big(\frac{1}{n}\Big).$$

In the rest of the paper, we assume that

$$0 \le p \le \Big(\frac{3\log n}{m}\Big)^{1/2}.$$

As described in the introduction, the random intersection graph $G(n,m;p)$ may be constructed using an $n \times m$ random matrix $R(n,m;p)$ whose rows are indexed by $v \in V$ and columns are indexed by $a \in M$. For fixed $a \in M$, the probability of $V_a := \{v \in V : a \in L_v\}$ being a fixed k-subset of V is $p^k(1-p)^{n-k}$ for integer $k \ge 2$. Hence V_a is the k-subset for some $a \in M$ with probability $1-(1-p^k(1-p)^{n-k})^m$, which will be approximated by

$$p_k := 1 - e^{-mp^k(1-p)^{n-k}}.$$

Also, $G(n,m;p)$ will be approximated by another random graph $G(n,(p_k))$, which is to be defined below.

For $0 \le p^* \le 1$, let $\mathcal{H}_k(n,p^*)$ be a random collection of k-subsets of V to which each k-subset belongs with probability p^*, independently of all other k-subsets. For $H \subseteq V$, let $K(H)$ be the complete graph on H. Then, for a collection $\mathcal{H}$ of subsets of V, let $K(\mathcal{H})$ denote the graph on V whose edge set is the union of edge sets of the complete graphs $K(H)$ on $H \in \mathcal{H}$. Notice that $K(\mathcal{H}_2(n,p^*))$ is the binomial random graph $G(n,p^*)$. For p_k defined above, let $G(n,(p_k))$ be the random graph on V whose edge set is the union of edge sets of $K(\mathcal{H}_2(n,p_2)), K(\mathcal{H}_3(n,p_3)), \ldots, K(\mathcal{H}_k(n,p_k)), \ldots$.

For $m \gg n^4$ and $p \le \left(\frac{3\log n}{m}\right)^{1/2}$, the probability of $\bigcup_{k\ge5} \mathcal{H}_k(n,p_k)$ being nonempty is upper bounded by

$$\sum_{k\ge5} \binom{n}{k} p_k \le \sum_{k\ge5} n^k m p^k = O\Big(\frac{n^5\log^3 n}{m^{3/2}}\Big) = O\Big(\frac{\log^3 n}{n}\Big).$$

Thus, for $G(n,p_2,p_3,p_4) = G(n,(p_2,p_3,p_4,\ 0,\ldots))$,

$$\mathrm{TV}\Big(G(n,(p_k)), G(n,p_2,p_3,p_4)\Big) \le \Pr\Big[\bigcup_{k\ge5} \mathcal{H}_k(n,p_k) \ne \emptyset\Big] = O\Big(\frac{\log^3 n}{n}\Big).$$

We will further approximate $G(n,p_2,p_3,p_4)$ by $G(n,p_2)$, which is the main contribution of this paper.

Summarizing all, since the total variation distance between $G(n,m;p)$ and $G(n,\hat{p})$ is upper bounded by the sum of $\mathrm{TV}(G(n,m;p),G(n,(p_k)))$, $\mathrm{TV}(G(n,(p_k)),G(n,p_2,p_3,p_4))$, $\mathrm{TV}(G(n,p_2,p_3,p_4),G(n,p_2))$ and $\mathrm{TV}(G(n,p_2),G(n,\hat{p}))$, it is enough to show that each total variation distance tends to 0. For the second one is $O\Big(\frac{\log^3 n}{n}\Big)$ described as above, we will prove that the other three total variation distances tend to 0 in Sections 3, 4 and 5, respectively.

3. Total variation distance between $G(n,m;p)$ and $G(n,(p_k))$

To prove that the total variation distance between $G(n,m;p)$ and $G(n,(p_k))$ tends to 0, we will use a coupling argument. For two random variables X and Y, a coupling (X',Y') of X and Y is a vector of random variables such that the marginal distributions of (X',Y') are the distributions of X and Y, respectively. The total variation distance between X and Y is upper bounded by the probability of $X'\neq Y'$ for any coupling (X',Y') of X and Y. On the other hand, there always exists a coupling (X',Y') so that the total variation distance of X and Y is equal to the probability of $X'\neq Y'$.

Lemma 3.1. [22, Chapter I, Theorem 5.2] *Let X and Y be random variables. Then any coupling (X',Y') of X and Y satisfies*

$$\mathrm{TV}(X,Y)\leq \Pr[X'\neq Y'].$$

Moreover, there exists a coupling for which the equality holds, i.e.,

$$\mathrm{TV}(X,Y)= \Pr[X'\neq Y'].$$

Using an appropriate coupling between a binomial random variable and a Poisson random variable, we will prove the following proposition, which may be applied for the case $m\gg n^2\log n$. The proposition is essentially the same as Lemma 5 in [26]. We prove it for the sake of completeness.

Proposition 3.2. *Let $m\gg n^2\log n$, $0\leq p\leq (\frac{3\log n}{m})^{1/2}$ and $p_k=1-e^{-mp^k(1-p)^{n-k}}$ for integers $k\geq 2$. Then*

$$\mathrm{TV}\Big(G(n,m;p),G(n,(p_k))\Big)=O\Big(\frac{n^2\log n}{m}\Big).$$

Proof. Let X be the number of columns of the matrix $R(n,m;p)$ with two or more 1's, or equivalently, the number of $a\in M$ with $|V_a|\geq 2$. Since

$$\Pr[|V_a|=k]=\binom{n}{k}p^k(1-p)^{n-k}=:r_k$$

for any fixed $a\in M$, the random variable X has the binomial distribution with parameters m and $q_2:=\sum_{k\geq 2}r_k$, i.e.,

$$\Pr[X=\ell]=\binom{m}{\ell}(q_2)^\ell(1-q_2)^{m-\ell}.$$

The random graph $G(n,m;p)$ may be constructed as follows: Take i.i.d. random complete graphs $K^{(1)},...,K^{(h)},...$ on subsets of V, where the number of vertices in $K^{(1)}$ is $k\geq 2$ with probability r_k/q_2, and then, once the number is given to be k, every k-subset of V is equally likely to be the vertex set of $K^{(1)}$. In other words, for a k-subset U of V with $k\geq 2$, the probability of U being the vertex set of $K^{(1)}$ is $\frac{r_k}{q_2}\binom{n}{k}^{-1}$. (As $\sum_{k\geq 2}\frac{r_k}{q_2}=1$, the random complete graph $K^{(1)}$ is well-defined.) The edge set of $G(n,m;p)$ is the union of edge sets of X random complete graphs $K^{(1)},...,K^{(X)}$.

We now take a Poisson random variable Y with mean mq_2 that is coupled with X so that $\Pr[X\neq Y]=\mathrm{TV}(X,Y)$. Let G_Y be the graph whose edge set is the union of edge sets of $K^{(1)},...,K^{(Y)}$. Then

$$\mathrm{TV}(G(n,m;p),G_Y)\leq \Pr[G(n,m;p)\neq G_Y]\leq \Pr[X\neq Y]=\mathrm{TV}(X,Y).$$

On the other hand, G_Y has the same distribution as $G(n,(p_k))$. Indeed, for each subset U of V with $|U|\geq 2$, let $Z(U)$ be the number of $i=1,2,...,Y$ such that the vertex set of $K^{(i)}$ is U. Then, it is well-known that for $k=|U|$, $Z(U)$'s are independent Poisson random variables with mean

$mq_2 \cdot \frac{r_k}{q_2}\binom{n}{k}^{-1} = mp^k(1-p)^{n-k}$, and hence $\Pr[Z(U) > 0] = 1 - e^{-mp^k(1-p)^{n-k}} = p_k$. Since the edge set of G_Y is the union of edge sets of the complete graphs on U with $Z(U) > 0$, G_Y has the same distribution as $G(n, (p_k))$.

The desired bound follows from the fact that the total variation distance between the binomial random variable X with parameters m, q_2 and the Poisson random variable Y with mean mq_2 is not more than q_2 [1, Theorem 2.4], and

$$q_2 = \sum_{k\geq 2}\binom{n}{k}p^k(1-p)^{n-k} \leq \sum_{k\geq 2} n^k p^k = O\big(n^2p^2\big) = O\Big(\frac{n^2\log n}{m}\Big).$$

□

4. TOTAL VARIATION DISTANCE BETWEEN $G(n, p_2, p_3, p_4)$ AND $G(n, p_2)$

In this section, we prove that the total variation distance between $G(n, p_2, p_3, p_4)$ and $G(n, p_2)$ tends to 0. This is the main contribution of the paper. Intuitively, if there are no artifact triangles (and no columns with at least four 1's) with high probability, then $G(n, p_2, p_3, p_4)$ and $G(n, p_2)$ should be almost the same. We will show that $\mathrm{TV}(G(n, p_2, p_3, p_4), G(n, p_2))$ is still small enough even if there are few artifact triangles. As mentioned earlier, it actually turns out that the distance is small enough if the expected number of pairs of distinct artifact triangles with a common edge is small. When the expected number is not small, the total variation distance tends to 0 provided that mp^2 is sufficiently large. Keeping this in mind, we prove the following proposition.

Proposition 4.1. *Let $m \gg n^4$, $0 \leq p \leq (\frac{3\log n}{m})^{1/2}$ and $p_k = 1 - e^{-mp^k(1-p)^{n-k}}$ for $k \geq 2$. Then*

$$\mathrm{TV}\Big(G(n, p_2, p_3, p_4), G(n, p_2)\Big) = O(\varepsilon),$$

where

$$\varepsilon := \max\Big\{\frac{1}{\log n}, \frac{1}{\log(m/n^4)}\Big\}. \tag{1}$$

For simplicity, we write $G(n, \mathbf{p}_4)$ for $G(n, p_2, p_3, p_4)$. It is not difficult to check that

$$\mathrm{TV}\Big(G(n,\mathbf{p}_4), G(n,p_2)\Big) = \sum_{G\in\mathcal{G}}\Big(\Pr[G(n,p_2) = G] - \min\big\{\Pr[G(n,\mathbf{p}_4) = G], \Pr[G(n,p_2) = G]\big\}\Big), \tag{2}$$

where $\mathcal{G}$ is the set of all graphs on V. In order to bound the total variation distance, we consider a lower bound of $\Pr[G(n, \mathbf{p}_4) = G]$. Since $G(n, \mathbf{p}_4) = K(\mathcal{H}_4(n, p_4)) \cup K(\mathcal{H}_3(n, p_3)) \cup G(n, p_2)$, we may write $\Pr[G(n, \mathbf{p}_4) = G]$ as the sum of

$$\Pr\Big[\mathcal{H}_4(n,p_4) = Q,\ \ \mathcal{H}_3(n,p_3) = T,\ \ G\setminus(K(T)\cup K(Q)) \subseteq G(n,p_2) \subseteq G\Big] \tag{3}$$

over all possible T and Q. Let $\mathcal{H}_3(G)$ and $\mathcal{H}_4(G)$ be the collections of all K_3's and K_4's in G that are regarded as collections of 3-subsets and 4-subsets of V, respectively. Then,

$$\begin{aligned}\Pr[G(n,\mathbf{p}_4) = G] &= \sum_{\substack{Q\subseteq\mathcal{H}_4(G)\\ T\subseteq\mathcal{H}_3(G)}} p_4^{|Q|}(1-p_4)^{\binom{n}{4}-|Q|}p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|}p_2^{|G|-|K(Q)\cup K(T)|}(1-p_2)^{\binom{n}{2}-|G|}\\ &= \Pr[G(n,p_2) = G]\sum_{\substack{Q\subseteq\mathcal{H}_4(G)\\ T\subseteq\mathcal{H}_3(G)}} p_4^{|Q|}(1-p_4)^{\binom{n}{4}-|Q|}p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|}p_2^{-|K(Q)\cup K(T)|},\end{aligned}$$

where $|G|$ is the number of edges in G. Let $G\setminus K(Q)$ be the graph obtained from G by removing the edges of the graph $K(Q)$. For each $Q \subseteq \mathcal{H}_4(G)$, taking only the case that $T \subseteq \mathcal{H}_3(G\setminus K(Q))$ yields that

$$\frac{\Pr[G(n,\mathbf{p}_4) = G]}{\Pr[G(n,p_2) = G]} \geq \sum_{Q\subseteq\mathcal{H}_4(G)} p_4^{|Q|}(1-p_4)^{\binom{n}{4}-|Q|}p_2^{-|K(Q)|}\sum_{T\subseteq\mathcal{H}_3(G\setminus K(Q))} p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|}p_2^{-|K(T)|}. \tag{4}$$

In the case that the expected number $\binom{n}{3}p_3 = \Theta(n^3mp^3)$ of artifact triangles is small, say $p \le \frac{\varepsilon}{nm^{1/3}}$, one may take $T, Q = \emptyset$ in the lower bound of (4) to obtain

$$\Pr[G(n,\mathbf{p}_4) = G] \ge \Pr[G(n,p_2) = G](1-p_4)^{\binom{n}{4}}(1-p_3)^{\binom{n}{3}},$$

and then (2) gives that

$$\mathrm{TV}\Big(G(n,\mathbf{p}_4), G(n,p_2)\Big) \le \sum_{G\in\mathcal{G}} \Pr[G(n,p_2) = G]\left(1-(1-p_4)^{\binom{n}{4}}(1-p_3)^{\binom{n}{3}}\right) = O(\varepsilon)$$

as $\binom{n}{3}p_3 = \Theta(n^3mp^3) = O(\varepsilon)$ and $\binom{n}{4}p_4 = \Theta(n^4mp^4) = O(\varepsilon)$. If $m = n^\alpha$ for $\alpha > 6$, then this holds for all $p \le \left(\frac{3\log n}{m}\right)^{1/2}$ since $\frac{\varepsilon}{nm^{1/3}} \ge \left(\frac{3\log n}{m}\right)^{1/2}$, which essentially implies the result of [13].

We now assume that

$$\frac{\varepsilon}{nm^{1/3}} < p \le \left(\frac{3\log n}{m}\right)^{1/2}.$$

For any set $\mathcal{G}^*$ of graphs on V, using (2), we have that the total variation distance is at most

$$\Pr[G(n,p_2) \notin \mathcal{G}^*] + \sum_{G\in\mathcal{G}^*}\Big(\Pr[G(n,p_2) = G] - \min\big\{\Pr[G(n,\mathbf{p}_4) = G], \Pr[G(n,p_2) = G]\big\}\Big).$$

Therefore it should be enough to consider the graphs G satisfying

$$|\mathcal{H}_3(G)| \approx \binom{n}{3}p_2^3 \text{ and } |\mathcal{H}_4(G)| \approx \binom{n}{4}p_2^6,$$

the exact meaning of which will be defined later.

We first give an intuition behind the proof that will be given later. Recalling (4), it turns out that

$$\begin{aligned}\sum_{T\subseteq\mathcal{H}_3(G\setminus K(Q))} p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|}p_2^{-|K(T)|} &\le \sum_{t\ge 0}\sum_{\substack{T\subseteq\mathcal{H}_3(G)\\|T|=t}} p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-|K(T)|}\\ &\lesssim \sum_{t\ge0}\binom{\binom{n}{3}p_2^3}{t}p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-3t}. \qquad (5)\end{aligned}$$

Since $\binom{\binom{n}{3}p_2^3}{t} \le \binom{\binom{n}{3}}{t}p_2^{3t}$, it follows that

$$\sum_{t\ge0}\binom{\binom{n}{3}p_2^3}{t}p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-3t} \le \sum_{t\ge0}\binom{\binom{n}{3}}{t}p_3^t(1-p_3)^{\binom{n}{3}-t} = 1.$$

Similarly,

$$\sum_{Q\subseteq\mathcal{H}_4(G)} p_4^{|Q|}(1-p_4)^{\binom{n}{4}-|Q|}p_2^{-|K(Q)|} \lesssim 1.$$

Therefore, the lower bound of (4) is close to 1 only when all the upper bounds are quite tight. In particular, to have the inequality (5) tight, we need that $|K(T)| = 3t$ for most collections T of t triangles in G for $t \approx \binom{n}{3}p_3$ unless p_2 is almost 1. If t is not close to $\binom{n}{3}p_3$, then the summands are small enough to be negligible. Note that $|K(T)| = 3t$ means that there is no pair of triangles in $\mathcal{H}_3(n,p_3) = T$ with a common edge. We consider two cases below depending upon whether the expected number of pairs of artifact triangles $\Theta(\binom{n}{4}\binom{m}{2}p^6) = \Theta(n^4m^2p^6)$ is small or not.

We will prove the following two lemmas, from which the main proposition easily follows. Recall that $\varepsilon = \max\left\{\frac{1}{\log n}, \frac{1}{\log(m/n^4)}\right\}$ and $p_k = 1 - e^{-mp^k(1-p)^{n-k}}$.

Lemma 4.2. *Suppose that*

$$m \gg n^4 \quad \textit{and} \quad \frac{\varepsilon}{nm^{1/3}} < p \leq \frac{\varepsilon}{n^{2/3}m^{1/3}}.$$

Then

$$\mathrm{TV}\Big(G(n, p_2, p_3, p_4), G(n, p_2)\Big) = O(\varepsilon).$$

Lemma 4.3. *Suppose that*

$$m \gg n^4 \quad \textit{and} \quad \frac{\varepsilon}{n^{2/3}m^{1/3}} < p \leq \Big(\frac{3\log n}{m}\Big)^{1/2}.$$

Then

$$\mathrm{TV}\Big(G(n, p_2, p_3, p_4), G(n, p_2)\Big) = O(\varepsilon).$$

(If m/n^4 is too large, e.g., $m = n^5$, then there is no such p, so the conclusion is trivially true. On the other hand, if it is not too large, e.g., $m = n^4 \log\log n$, then there are p satisfying the conditions.)

Before we prove Lemmas 4.2 and 4.3, three preliminary lemmas are introduced.

Lemma 4.4. *For $m \gg n^4$ and $\frac{\varepsilon}{nm^{1/3}} < p \leq \frac{\varepsilon}{n^{2/3}m^{1/3}}$, suppose that a graph G on V satisfies*

(i) $|\mathcal{H}_3(G)| \geq (1-\delta)\binom{n}{3}p_2^3$, where $\delta := \frac{1}{\varepsilon}\Big(\frac{1-p_2}{n^2p_2} + \frac{1-p_2}{n^3p_2^3}\Big)^{1/2}$,

(ii) the number $I(G)$ of diamond graphs (i.e., K_4 minus one edge) in G is at most $n^4p_2^5/\varepsilon$.

Then the number of sets T such that $T \subseteq \mathcal{H}_3(G)$, $|T| = t$ and $|K(T)| = 3t$ is at least

$$(1 - O(\varepsilon))\binom{\binom{n}{3}}{t}p_2^{3t} \qquad \textit{for } \ 0 \leq t \leq t_0 := \frac{n^3mp^3}{\varepsilon},$$

where the constant in $O(\varepsilon)$ is independent of G and t.

Proof. Let $X_t(G)$ be the number of sets T such that $T \subseteq \mathcal{H}_3(G)$, $|T| = t$ and $|K(T)| = 3t$. We infer that

$$\begin{aligned} X_t(G) &\geq \binom{|\mathcal{H}_3(G)|}{t} - I(G)\binom{|\mathcal{H}_3(G)|}{t-2} \\ &= \Big(1 - \frac{t(t-1)I(G)}{(|\mathcal{H}_3(G)|-t+2)(|\mathcal{H}_3(G)|-t+1)}\Big)\binom{|\mathcal{H}_3(G)|}{t} \\ &\geq \Big(1 - \frac{t_0^2 I(G)}{(|\mathcal{H}_3(G)|-t_0)^2}\Big)\binom{|\mathcal{H}_3(G)|}{t}. \end{aligned}$$

Since $|\mathcal{H}_3(G)| = \Omega(n^3p_2^3)$, we have that

$$\frac{t_0^2}{|\mathcal{H}_3(G)|} = O\Big(\frac{n^3m^2p^6}{\varepsilon^2p_2^3}\Big) = O\Big(\frac{n^3}{\varepsilon^2 m} + \frac{n^3m^2p^6}{\varepsilon^2}\Big) = O\Big(\frac{\varepsilon^2}{n}\Big), \tag{6}$$

where the second equality follows from $p_2 = \Theta\Big(\frac{mp^2}{1+mp^2}\Big)$ and the third equality follows from $p \leq \frac{\varepsilon}{n^{2/3}m^{1/3}}$. In particular $|\mathcal{H}_3(G)| \gg t_0$, and hence

$$X_t(G) \geq \Big(1 - \frac{2t_0^2 I(G)}{|\mathcal{H}_3(G)|^2}\Big)\binom{|\mathcal{H}_3(G)|}{t}.$$

It is easy to check from (6) that

$$\frac{2t_0^2 I(G)}{|\mathcal{H}_3(G)|^2} = \frac{2I(G)}{|\mathcal{H}_3(G)|} \cdot \frac{t_0^2}{|\mathcal{H}_3(G)|} = O\Big(\frac{n^4p_2^5}{\varepsilon n^3 p_2^3} \cdot \frac{\varepsilon^2}{n}\Big) = O(\varepsilon)$$

and

$$\binom{|\mathcal{H}_3(G)|}{t} \geq \left(1 - \frac{t_0}{|\mathcal{H}_3(G)|}\right)^{t_0} \frac{|\mathcal{H}_3(G)|^t}{t!} \geq \left(1 - O\left(\frac{\varepsilon^2}{n}\right)\right) \frac{|\mathcal{H}_3(G)|^t}{t!} \geq (1 - O(\varepsilon))(1-\delta)^{t_0} \binom{\binom{n}{3}}{t} p_2^{3t}$$

as $|\mathcal{H}_3(G)| \geq (1-\delta)\binom{n}{3}p_2^3$. Since $p_2 = \Theta\left(\frac{mp^2}{1+mp^2}\right)$ and $p \leq \frac{\varepsilon}{n^{2/3}m^{1/3}}$ yield

$$(\delta t_0)^2 = O\left(\frac{n^4 mp^4}{\varepsilon^4} + \frac{n^3}{\varepsilon^4 m}\right) = O\left(\left(\frac{n^4}{m}\right)^{1/3} + \frac{1}{n}\left(\frac{n^4}{\varepsilon^4 m}\right)\right) = O(\varepsilon^2), \tag{7}$$

the desired lower bound for $X_t(G)$ follows. □

The same argument gives the next lemma regarding $\mathcal{H}_4(G)$.

Lemma 4.5. *For $m \gg n^4$ and $\frac{\varepsilon}{n^{2/3}m^{1/3}} < p \leq \left(\frac{3\log n}{m}\right)^{1/2}$, suppose that a graph G on V satisfies*

$$|\mathcal{H}_4(G)| \geq \left(1 - \frac{1}{\varepsilon n}\right)\binom{n}{4}p_2^6.$$

Then the number of $Q \subseteq \mathcal{H}_4(G)$ with $|Q| = q$ and $|K(Q)| = 6q$ is at least

$$(1 - O(\varepsilon))\binom{\binom{n}{4}}{q} p_2^{6q} \qquad \text{for } \ 0 \leq q \leq q_0 := \frac{n^4 mp^4}{\varepsilon},$$

where the constant in $O(\varepsilon)$ is independent of G and t.

Remark. The expected number of columns of the matrix $R(n,m;p)$ with four or more 1's is $\Theta(n^4mp^4)$. The parameter q_0 is chosen to be substantially, but not extremely, bigger than the expected number $\Theta(n^4mp^4)$.

Proof. Let $Y_q(G)$ be the number of $Q \subseteq \mathcal{H}_4(G)$ with $|Q| = q$ and $|K(Q)| = 6q$. Observe that the number of pairs of K_4 in the complete graph on V sharing at least an edge is at most $\binom{n}{4}\binom{n}{2}\binom{4}{2} \leq n^6$. Thus

$$\begin{aligned} Y_q(G) &\geq \binom{|\mathcal{H}_4(G)|}{q} - n^6\binom{|\mathcal{H}_4(G)|}{q-2} \\ &= \left(1 - \frac{q(q-1)n^6}{(|\mathcal{H}_4(G)| - q + 2)(|\mathcal{H}_4(G)| - q + 1)}\right)\binom{|\mathcal{H}_4(G)|}{q} \\ &\geq \left(1 - \frac{q_0^2 n^6}{(|\mathcal{H}_4(G)| - q_0)^2}\right)\binom{|\mathcal{H}_4(G)|}{q}. \end{aligned}$$

Since $mp^2 \geq m\left(\frac{\varepsilon}{n^{2/3}m^{1/3}}\right)^2 = \varepsilon^2\left(\frac{m}{n^4}\right)^{1/3} \to \infty$, we have that $p_2 = 1 - o(1)$ and

$$|\mathcal{H}_4(G)| \geq \left(1 - \frac{1}{\varepsilon n}\right)\binom{n}{4}p_2^6 = (1 - o(1))\binom{n}{4}.$$

Therefore $q_0 = \frac{n^4mp^4}{\varepsilon} = O(\varepsilon \log^2 n)$ implies that

$$\frac{q_0^2 n^6}{(|\mathcal{H}_4(G)| - q_0)^2} = O(\varepsilon),$$

and hence

$$Y_q(G) \geq (1 - O(\varepsilon))\binom{|\mathcal{H}_4(G)|}{q}.$$

Since $\frac{q_0^2}{|\mathcal{H}_4(G)|} = O\left(\frac{\varepsilon^2 \log^4 n}{n^4}\right) = O(\varepsilon)$ and $\frac{q_0}{\varepsilon n} = O\left(\frac{\log^2 n}{n}\right) = O(\varepsilon)$, we have that

$$\binom{|\mathcal{H}_4(G)|}{q} \geq \left(1 - \frac{q_0}{|\mathcal{H}_4(G)|}\right)^{q_0}\frac{|\mathcal{H}_4(G)|^q}{q!} \geq (1 - O(\varepsilon))\left(1 - \frac{1}{\varepsilon n}\right)^{q_0}\binom{\binom{n}{4}}{q}p_2^{6q} \geq (1 - O(\varepsilon))\binom{\binom{n}{4}}{q}p_2^{6q},$$

which gives the desired lower bound for $Y_q(G)$. $\square$

Lemma 4.6. *For $\delta = \frac{1}{\varepsilon}\left(\frac{1-p_2}{n^2 p_2} + \frac{1-p_2}{n^3 p_2^3}\right)^{1/2}$, let $\mathcal{G}_3$ be the set of all graphs G on V satisfying*

$$|\mathcal{H}_3(G)| \geq (1-\delta)\binom{n}{3}p_2^3 \text{ and } I(G) \leq n^4 p_2^5/\varepsilon,$$

recalling that $I(G)$ denotes the number of diamond graphs as in Lemma 4.4, and let $\mathcal{G}_4$ be the set of all graphs G in $\mathcal{G}_3$ satisfying

$$|\mathcal{H}_4(G)| \geq \left(1 - \frac{1}{\varepsilon n}\right)\binom{n}{4}p_2^6.$$

Then for $m \gg n^4$ we have

$$\Pr[G(n,p_2) \in \mathcal{G}_3] = 1 - O(\varepsilon) \qquad \text{for } \frac{\varepsilon}{nm^{1/3}} < p \leq \left(\frac{3\log n}{m}\right)^{1/2}$$

and

$$\Pr[G(n,p_2) \in \mathcal{G}_4] = 1 - O(\varepsilon) \qquad \text{for } \frac{\varepsilon}{n^{2/3}m^{1/3}} < p \leq \left(\frac{3\log n}{m}\right)^{1/2},$$

where the constants in $O(\varepsilon)$ are independent of p.

Proof. For $X_3 := |\mathcal{H}_3(G(n,p_2))|$, Chebyshev's inequality gives that

$$\Pr\left[X_3 < (1-\delta)\binom{n}{3}p_2^3\right] \leq \Pr\left[|X_3 - E[X_3]| > \delta\binom{n}{3}p_2^3\right] \leq \frac{\mathrm{Var}[X_3]}{\delta^2\binom{n}{3}^2 p_2^6} = O(\varepsilon^2)$$

as $E[X_3] = \binom{n}{3}p_2^3$ and $\mathrm{Var}[X_3] = O\left((n^4 p_2^5 + n^3 p_2^3)(1-p_2)\right)$. Moreover, Markov's inequality implies that

$$\Pr\left[I(G(n,p_2)) > \frac{n^4 p_2^5}{\varepsilon}\right] \leq \varepsilon$$

since $E\left[I(G(n,p_2))\right] = \binom{n}{4}6 \cdot p_2^5 \leq n^4 p_2^5$. Therefore,

$$\Pr[G(n,p_2) \notin \mathcal{G}_3] \leq \Pr\left[X_3 < (1-\delta)\binom{n}{3}p_2^3\right] + \Pr\left[I(G(n,p_2)) > \frac{n^4 p_2^5}{\varepsilon}\right] = O(\varepsilon).$$

Similarly, for $X_4 = |\mathcal{H}_4(G(n,p_2))|$, it is not hard to see that

$$E[X_4] = \binom{n}{4}p_2^6 \text{ and } \mathrm{Var}[X_4] = O\left(n^6\right)$$

as $p_2 = 1 - o(1)$ for $p > \frac{\varepsilon}{n^{2/3}m^{1/3}}$, and Chebyshev's inequality yields that

$$\Pr\left[X_4 < \left(1 - \frac{1}{\varepsilon n}\right)\binom{n}{4}p_2^6\right] \leq \Pr\left[|X_4 - E[X_4]| > \frac{1}{\varepsilon n}\binom{n}{4}p_2^6\right] \leq \frac{\varepsilon^2 n^2 \,\mathrm{Var}[X_4]}{\binom{n}{4}^2 p_2^{12}} = O(\varepsilon^2).$$

Therefore,

$$\Pr[G(n,p_2) \notin \mathcal{G}_4] \leq \Pr[G(n,p_2) \notin \mathcal{G}_3] + \Pr\left[X_4 < \left(1 - \frac{1}{\varepsilon n}\right)\binom{n}{4}p_2^6\right] = O(\varepsilon).$$

$\square$

Now we prove the main lemmas.

Proof of Lemma 4.2. Equality (2) and Lemma 4.6 imply that the total variation distance between $G(n, \mathbf{p}_4)$ and $G(n, p_2)$ is at most

$$\begin{aligned}
&\Pr[G(n,p_2) \notin \mathcal{G}_3] + \sum_{G \in \mathcal{G}_3} \Big(\Pr[G(n,p_2) = G] - \min\big\{ \Pr[G(n,\mathbf{p}_4) = G], \Pr[G(n,p_2) = G] \big\} \Big) \\
&= O(\varepsilon) + \sum_{G \in \mathcal{G}_3} \Big(\Pr[G(n,p_2) = G] - \min\big\{ \Pr[G(n,\mathbf{p}_4) = G], \Pr[G(n,p_2) = G] \big\} \Big).
\end{aligned} \tag{8}$$

Taking $Q = \emptyset$ in (4), we have that

$$\begin{aligned}
\Pr[G(n,\mathbf{p}_4) = G] &\geq \Pr[G(n,p_2) = G](1-p_4)^{\binom{n}{4}} \sum_{T \subseteq \mathcal{H}_3(G)} p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|} p_2^{-|K(T)|} \\
&= (1 - O(\varepsilon)) \Pr[G(n,p_2) = G] \sum_{T \subseteq \mathcal{H}_3(G)} p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|} p_2^{-|K(T)|}
\end{aligned}$$

as $\binom{n}{4} p_4 = \Theta(n^4 m p^4) = O(\varepsilon)$. For $G \in \mathcal{G}_3$, Lemma 4.4 gives that

$$\begin{aligned}
\sum_{T \subseteq \mathcal{H}_3(G)} p_3^{|T|}(1-p_3)^{\binom{n}{3}-|T|} p_2^{-|K(T)|} &\geq \sum_{t=0}^{t_0} \sum_{\substack{T \subseteq \mathcal{H}_3(G) \\ |T|=t, |K(T)|=3t}} p_3^{t}(1-p_3)^{\binom{n}{3}-t} p_2^{-3t} \\
&\geq (1 - O(\varepsilon)) \sum_{t=0}^{t_0} \binom{\binom{n}{3}}{t} p_3^t (1-p_3)^{\binom{n}{3}-t},
\end{aligned}$$

and

$$\frac{\Pr[G(n,\mathbf{p}_4) = G]}{\Pr[G(n,p_2) = G]} \geq (1 - O(\varepsilon)) \sum_{t=0}^{t_0} \binom{\binom{n}{3}}{t} p_3^t (1-p_3)^{\binom{n}{3}-t}.$$

Since $t_0 = n^3 m p^3/\varepsilon = \Theta(n^3 p_3/\varepsilon)$, Markov's inequality yields that

$$\sum_{t=0}^{t_0} \binom{\binom{n}{3}}{t} p_3^t (1-p_3)^{\binom{n}{3}-t} = 1 - \Pr\left[\mathrm{Bin}\left(\binom{n}{3}, p_3 \right) > t_0 \right] = 1 - O(\varepsilon),$$

where $\mathrm{Bin}(n', p')$ is the binomial random variable with parameters n' and p'. Therefore,

$$\Pr[G(n,\mathbf{p}_4) = G] \geq (1 - O(\varepsilon)) \Pr[G(n,p_2) = G] \quad \text{for } G \in \mathcal{G}_3,$$

which together with (8) implies that $\mathrm{TV}\Big(G(n,\mathbf{p}_4), G(n,p_2)\Big) = O(\varepsilon)$, provided

$$m \gg n^4 \quad \text{and} \quad \frac{\varepsilon}{n^{2/3} m^{1/3}} < p \leq \left(\frac{3 \log n}{m} \right)^{1/2}.$$

□

Proof of Lemma 4.3. As in the proof of Lemma 4.2, it follows from (2) and Lemma 4.6 that

$$\begin{aligned}
&\mathrm{TV}\Big(G(n,\mathbf{p}_4), G(n,p_2)\Big) \\
\leq\ & O(\varepsilon) + \sum_{G \in \mathcal{G}_4} \Big(\Pr[G(n,p_2) = G] - \min\big\{ \Pr[G(n,\mathbf{p}_4) = G], \Pr[G(n,p_2) = G] \big\} \Big).
\end{aligned} \tag{9}$$

Let $Q \subseteq \mathcal{H}_4(G)$, and we write $G \setminus Q$ for $G \setminus K(Q)$ for brevity. For $G \in \mathcal{G}_4$, the sum in the lower bound of (4) restricted to the cases $|T| \le t_0 = n^3mp^3/\varepsilon$ and $|Q| \le q_0 = n^4mp^4/\varepsilon$ gives

$$\frac{\Pr[G(n,\mathbf{p}_4)=G]}{\Pr[G(n,p_2)=G]} \ge \sum_{q=0}^{q_0} \sum_{\substack{Q\subseteq\mathcal{H}_4(G)\\|Q|=q,|K(Q)|=6q}} p_4^q(1-p_4)^{\binom{n}{4}-q}p_2^{-6q} \sum_{t=0}^{t_0} \sum_{\substack{T\subseteq\mathcal{H}_3(G\setminus Q)\\|T|=t}} p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-|K(T)|}.$$

Lemma 4.5 and Markov's inequality imply that

$$\begin{aligned}\sum_{q=0}^{q_0} \sum_{\substack{Q\subseteq\mathcal{H}_4(G)\\|Q|=q,|K(Q)|=6q}} p_4^q(1-p_4)^{\binom{n}{4}-q}p_2^{-6q} &\ge (1-O(\varepsilon))\sum_{q=0}^{q_0}\binom{\binom{n}{4}}{q}p_4^q(1-p_4)^{\binom{n}{4}-q}\\ &= (1-O(\varepsilon))\left(1-\Pr\left[\mathrm{Bin}\left(\binom{n}{4},p_4\right)>q_0\right]\right)\\ &= 1-O(\varepsilon),\end{aligned}$$

where $\mathrm{Bin}(n',p')$ is a binomial random variable with parameters n' and p'. Therefore,

$$\frac{\Pr[G(n,\mathbf{p}_4)=G]}{\Pr[G(n,p_2)=G]} \ge (1-O(\varepsilon)) \min_{\substack{Q\subseteq\mathcal{H}_4(G)\\|Q|\le q_0}} \sum_{t=0}^{t_0} \sum_{\substack{T\subseteq\mathcal{H}_3(G\setminus Q)\\|T|=t}} p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-|K(T)|}. \tag{10}$$

For $T \subseteq \mathcal{H}_3(G)$, let $I^*(T)$ be the number of pairs of distinct triangles in T with a common edge. (It is a bit different from the definition $I(\cdot)$ in Lemma 4.4.) For an edge e, let $d_T(e)$ be the number of triangles in T which contain e. Then

$$3|T|-|K(T)| = \sum_{e:d_T(e)\ge 2}(d_T(e)-1) \le \sum_{e:d_T(e)\ge 2}\binom{d_T(e)}{2} = I^*(T).$$

For a fixed $Q \subseteq \mathcal{H}_4(G)$ with $|Q| \le q_0$, we will show that the number of $T \subseteq \mathcal{H}_3(G \setminus Q)$ with $|T| = t \le t_0 = n^3mp^3/\varepsilon$ and $I^*(T) \le r := n^4m^2p^6/\varepsilon^3$ is at least

$$(1-O(\varepsilon))\binom{\binom{n}{3}}{t}p_2^{3t}. \tag{11}$$

Then

$$\begin{aligned}\sum_{t=0}^{t_0} \sum_{\substack{T\subseteq\mathcal{H}_3(G\setminus Q)\\|T|=t}} p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-|K(T)|} &\ge \sum_{t=0}^{t_0} \sum_{\substack{T\subseteq\mathcal{H}_3(G\setminus Q)\\|T|=t,I^*(T)\le r}} p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-|K(T)|}\\ &\ge (1-O(\varepsilon))p_2^r\cdot\sum_{t=0}^{t_0}\binom{\binom{n}{3}}{t}p_3^t(1-p_3)^{\binom{n}{3}-t}\\ &\ge (1-O(\varepsilon))p_2^r,\end{aligned}$$

where the last inequality follows from Markov's inequality. Since $p_2^r = (1-e^{-mp^2(1-p)^{n-2}})^r \ge 1 - O(re^{-mp^2})$ and

$$re^{-mp^2} = \frac{n^4m^2p^6}{\varepsilon^3}e^{-mp^2} = \frac{n^4}{\varepsilon^3 m}\cdot(mp^2)^3e^{-mp^2} = O(\varepsilon),$$

we have that $p_2^r = 1-O(\varepsilon)$ and

$$\sum_{t=0}^{t_0} \sum_{\substack{T\subseteq\mathcal{H}_3(G\setminus Q)\\|T|=t}} p_3^t(1-p_3)^{\binom{n}{3}-t}p_2^{-|K(T)|} \ge 1-O(\varepsilon).$$

This together with (10) and (9) completes the proof of Lemma 4.3.

It remains to prove (11). For $t \le t_0$, we take the uniform random collection $R = R(t)$ of triangles that is equally likely to be T for every $T \subseteq \mathcal{H}_3(G \setminus Q)$ with $|T| = t$. In other words, for every $T \subseteq \mathcal{H}_3(G \setminus Q)$ with $|T| = t$,

$$\Pr[R = T] = \binom{|\mathcal{H}_3(G \setminus Q)|}{t}^{-1}.$$

Since the number of sets $T \subseteq \mathcal{H}_3(G \setminus Q)$ with $|T| = t$ containing a diamond graph is less than or equal to $I(G)\binom{|\mathcal{H}_3(G\setminus Q)|}{t-2}$, we have that

$$E[I^*(R)] \le I(G)\binom{|\mathcal{H}_3(G \setminus Q)|}{t-2}\binom{|\mathcal{H}_3(G \setminus Q)|}{t}^{-1} \le \frac{I(G)t_0^2}{(|\mathcal{H}_3(G \setminus Q)| - t_0)^2},$$

where $I(G)$ is defined in Lemma 4.4. For $G \in \mathcal{G}_4$, since $K(Q)$ has at most $6|Q| \le 6q_0 = \frac{6n^4mp^4}{\varepsilon} = O(\varepsilon \log^2 n)$ edges and each edge in G is contained in at most n triangles in $\mathcal{H}_3(G)$,

$$|\mathcal{H}_3(G \setminus Q)| \ge |\mathcal{H}_3(G)| - 6q_0 n = |\mathcal{H}_3(G)| - O(\varepsilon n \log^2 n) = \Theta(n^3). \tag{12}$$

As $t_0 = \frac{n^3mp^3}{\varepsilon} \ll n^3$ and $I(G) \le I(K_n) = 6\binom{n}{4} \le n^4$,

$$E[I^*(R)] = O\Big(\frac{I(G)t_0^2}{n^6}\Big) = O\Big(\frac{t_0^2}{n^2}\Big) = O\Big(\frac{n^4m^2p^6}{\varepsilon^2}\Big)$$

and Markov's inequality gives that

$$\Pr\left[I^*(R) > r\right] \le \frac{\varepsilon^3 E[I^*(R)]}{n^4m^2p^6} = O(\varepsilon).$$

The number Z of $T \subseteq \mathcal{H}_3(G \setminus Q)$ with $|T| = t$ and $I^*(T) \le r$ satisfies

$$Z = (1 - O(\varepsilon))\binom{|\mathcal{H}_3(G \setminus Q)|}{t}.$$

Now we estimate $\binom{|\mathcal{H}_3(G\setminus Q)|}{t}$. Since $G \in \mathcal{G}_4$ and $p_2 = 1 - o(1)$, it is obtained similarly to (12) that

$$|\mathcal{H}_3(G \setminus Q)| \ge |\mathcal{H}_3(G)| - 6q_0 n = \left(1 - \delta - O\left(\frac{q_0}{n^2}\right)\right)\binom{n}{3}p_2^3,$$

and then

$$\begin{aligned}\binom{|\mathcal{H}_3(G \setminus Q)|}{t} &\ge \Big(1 - \delta - O\Big(\frac{q_0}{n^2}\Big)\Big)^{t_0}\Big(1 - O\Big(\frac{t_0}{n^3}\Big)\Big)^{t_0}\frac{\binom{n}{3}^t}{t!}p_2^{3t}\\ &\ge \Big(1 - t_0\delta - O\Big(\frac{t_0 q_0}{n^2}\Big)\Big)\Big(1 - O\Big(\frac{t_0^2}{n^3}\Big)\Big)\binom{\binom{n}{3}}{t}p_2^{3t}.\end{aligned}$$

As in (7), $\delta t_0 = O(\varepsilon)$, and it is easy to check that

$$\frac{t_0 q_0}{n^2} = \frac{n^5m^2p^7}{\varepsilon^2} = O\Big(\frac{n^5\log^{7/2} n}{\varepsilon^2 m^{3/2}}\Big) = O(\varepsilon) \text{ and } \quad \frac{t_0^2}{n^3} = \frac{n^3m^2p^6}{\varepsilon^2} = O\Big(\frac{n^3\log^3 n}{\varepsilon^2 m}\Big) = O(\varepsilon).$$

Therefore, we have that

$$Z = (1 - O(\varepsilon))\binom{|\mathcal{H}_3(G \setminus Q)|}{t} \ge (1 - O(\varepsilon))\binom{\binom{n}{3}}{t}p_2^{3t}.$$

This completes the proof of (11). $\square$

5. TOTAL VARIATION DISTANCE BETWEEN $G(n, p_2)$ AND $G(n, \hat{p})$

For the random graphs $G(n,m;p), G(n,(p_k)), G(n,p_2,p_3,p_4)$ and $G(n,p_2)$, we have so far considered the total variation distance between the consecutive pairs of them. Finally, a good upper bound for the total variation distance between $G(n,p_2)$ and $G(n,\hat{p})$ easily follows from an upper bound for the total variation distance between two binomial distributions $\mathrm{Bin}(N,p)$ and $\mathrm{Bin}(N,q)$. As a corollary of Theorem 2.2 in [19], we may have

Corollary 5.1. *Let N be a positive integer, and p and q be real numbers satisfying $0 < p < q < 1$. For δ satisfying $(q-p)N = \delta\sqrt{p(1-p)N}$, i.e., $\delta = (q-p)\sqrt{\frac{N}{p(1-p)}}$, we have*

$$\mathrm{TV}\Big(\mathrm{Bin}(N,p), \mathrm{Bin}(N,q)\Big) \le \delta + 3\delta^2.$$

Recalling $p_2 = 1 - e^{-mp^2(1-p)^{n-2}}$, $\hat{p} = 1-(1-p^2)^m$ and $\varepsilon = \max\left\{\frac{1}{\log n}, \frac{1}{\log(m/n^4)}\right\}$, we have the last inequality needed.

Corollary 5.2. *Suppose that $m \gg n^4$ and $p \le \left(\frac{3\log n}{m}\right)^{1/2}$. Then*

$$\mathrm{TV}\Big(G(n,p_2), G(n,\hat{p})\Big) = O(\varepsilon).$$

Proof. Let $p = \sqrt{\frac{c}{m}}$ for $0 < c \le 3\log n$. Since $p_2 = \Theta\left(\frac{mp^2}{1+mp^2}\right) = \Theta\left(\frac{c}{1+c}\right)$ and

$$\hat{p} - p_2 = e^{-mp^2(1-p)^{n-2}} - e^{m\log(1-p^2)} \le e^{-mp^2}(e^{nmp^3} - e^{-mp^4}) = O(nmp^3e^{-mp^2}),$$

we have that

$$(\hat{p} - p_2)\sqrt{\frac{\binom{n}{2}}{p_2(1-p_2)}} = O\Big(\frac{n^2mp^3}{e^{mp^2}\sqrt{p_2(1-p_2)}}\Big) = O\Big(\Big(\frac{n^4}{m}\Big)^{1/2}\frac{c(1+c)}{e^c}\Big) = O(\varepsilon).$$

Therefore, Corollary 5.1 implies that

$$\mathrm{TV}\Big(G(n,p_2), G(n,\hat{p})\Big) \le \mathrm{TV}\Big(\mathrm{Bin}\Big(\binom{n}{2}, p_2\Big), \mathrm{Bin}\Big(\binom{n}{2}, \hat{p}\Big)\Big) = O(\varepsilon).$$

□

6. CONCLUDING REMARK

Fill, Scheinerman and Singer-Cohen [13] showed that the total variation distance between $G(n,m;p)$ and $G(n,\hat{p})$ tends to 0 for $m = n^\alpha, \alpha > 6$. In this paper, we improve the result. Namely, the total variation distance still goes to 0 for $m \gg n^4$. If $m \gg n^4$ then the expected number of pairs of artifact triangles with a common edge is small enough, or both of the two random graphs are complete graphs with high probability. This is the main ingredient of the proof of Theorem 1.2.

Our result naturally gives rise to the question whether the condition $m \gg n^4$ is tight. We initially believed that the total variation distance between $G(n,m;p)$ and $G(n,\hat{p})$ is not close to 0 if m is smaller than n^4. However, the more we try to prove it, the more we feel that our initial belief is baseless. It would not be extremely surprising even if the total variation distance tends to 0 for some m much less than n^4.

References

[1] A. D. Barbour and L. Holst. Some applications of the Stein-Chen method for proving Poisson convergence. *Adv. in Appl. Probab.*, 21(1):74–90, 1989.

[2] S. R. Blackburn and S. Gerke. Connectivity of the uniform random intersection graph. *Discrete Math.*, 309(16):5130–5140, 2009.

[3] M. Bloznelis. Degree distribution of a typical vertex in a general random intersection graph. *Lith. Math. J.*, 48(1):38–45, 2008.

[4] M. Bloznelis. Degree and clustering coefficient in sparse random intersection graphs. *Ann. Appl. Probab.*, 23(3):1254–1289, 2013.

[5] M. Bloznelis, E. Godehardt , J. Jaworski, V. Kurauskas and K, Rybarczyk. Recent Progress in Complex Network Analysis: Models of Random Intersection Graphs. Data Science, Learning by Latent Structures, and Knowledge Discovery. Part of the series Studies in Classification, Data Analysis, and Knowledge Organization, pp 69–78.

[6] M. Bloznelis, E. Godehardt , J. Jaworski, V. Kurauskas and K, Rybarczyk. Recent Progress in Complex Network Analysis: Properties of Random Intersection Graphs. Data Science, Learning by Latent Structures, and Knowledge Discovery. Part of the series Studies in Classification, Data Analysis, and Knowledge Organization, pp 79–88.

[7] M. Bloznelis, J. Jaworski, and V. Kurauskas. Assortativity and clustering of sparse random intersection graphs. *Electron. J. Probab.*, 18:no. 38, 24, 2013.

[8] M. Bloznelis, J. Jaworski, and K. Rybarczyk. Component evolution in a secure wireless sensor network. *Networks*, 53(1):19–26, 2009.

[9] T. Britton, M. Deijfen, A. N. Lagerås, and M. Lindholm. Epidemics on random graphs with tunable clustering. *J. Appl. Probab.*, 45(3):743–756, 2008.

[10] M. Deijfen and W. Kets. Random intersection graphs with tunable degree distribution and clustering. *Probab. Engrg. Inform. Sci.*, 23(4):661–674, 2009.

[11] R. Di Pietro, L. V. Mancini, A. Mei, A. Panconesi, and J. Radhakrishnan. Sensor networks that are provably resilient. *Securecomm and Workshops, 2006*, IEEE, 2006.

[12] P. Erdős, A. W. Goodman, and L. Pósa. The representation of a graph by set intersections. *Canad. J. Math.*, 18:106–112, 1966.

[13] J. A. Fill, E. R. Scheinerman, and K. B. Singer-Cohen. Random intersection graphs when $m = \omega(n)$: an equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models. *Random Structures Algorithms*, 16(2):156–176, 2000.

[14] F. Gavril. The intersection graphs of subtrees in trees are exactly the chordal graphs. *J. Combinatorial Theory Ser. B*, 16:47–56, 1974.

[15] E. Godehardt and J. Jaworski. Two models of random intersection graphs for classification. In *Exploratory data analysis in empirical research*, Stud. Classification Data Anal. Knowledge Organ., pages 67–81. Springer, Berlin, 2003.

[16] M. C. Golumbic. *Algorithmic graph theory and perfect graphs*, volume 57 of *Annals of Discrete Mathematics*. Elsevier Science B.V., Amsterdam, second edition, 2004.

[17] D. Gusfield. The multi-state perfect phylogeny problem with missing and removable data: Solutions via integer-programming and chordal graph theory. *Journal of Computational Biology*, 17(3):383–399, 2015/06/03 2010.

[18] L. Hubert. Some applications of graph theory and related non-metric techniques to problems of approximate seriation: The case of symmetric proximity measures. *British J. Math. Statist. Psychology*, 27(2):133–153, 1974.

[19] S. Janson. Asymptotic equivalence and contiguity of some random graphs. *Random Structures Algorithms*, 36(1):26–45, 2010.

[20] M. Karoński, E. R. Scheinerman, and K. B. Singer-Cohen. On random intersection graphs: the subgraph problem. *Combin. Probab. Comput.*, 8(1-2):131–159, 1999.

[21] H. Khamis and T. McKee. Chordal graph models of contingency tables. *Computers and Mathematics with Applications*, 34(11):89 – 97, 1997.

[22] T. Lindvall. *Lectures on the coupling method*. Dover Publications, Inc., Mineola, NY, 2002. Corrected reprint of the 1992 original.

[23] D. J. Marchette. *Random graphs for statistical pattern recognition*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2004.

[24] T. A. McKee and F. R. McMorris. *Topics in intersection graph theory*. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[25] K. Rybarczyk. Diameter, connectivity, and phase transition of the uniform random intersection graph. *Discrete Mathematics*, 311.17:1998–2019, 2011.

[26] K. Rybarczyk. Equivalence of a random intersection graph and $G(n, p)$. *Random Structures Algorithms*, 38(1-2):205–234, 2011.

[27] E. R. Scheinerman. Characterizing intersection classes of graphs. *Discrete Math.*, 55(2):185–193, 1985.

[28] P. G. Spirakis, S. Nikoletseas and C. Raptopoulos. A Guided Tour in Random Intersection Graphs. *Automata, Languages, and Programming.* Volume 7966 of the series Lecture Notes in Computer Science. pp 29-35.
[29] E. Szpilrajn-Marczewski. Sur deux propriétés des classes d'ensembles. *Fund. Math.*, 33:303–307, 1945.

School of Computational Sciences, Korea Institute for Advanced Study, Seoul, South Korea (J. H. Kim)
E-mail address: jhkim@kias.re.kr

Department of Mathematics, Duksung Women's University, Seoul, South Korea (S. J. Lee)
E-mail address: sanglee242@duksung.ac.kr, sjlee242@gmail.com

School of Computational Sciences, Korea Institute for Advanced Study, Seoul, South Korea (J. Na)
E-mail address: jhna@kias.re.kr, najoohan@gmail.com